

Graphical models for inferring single molecule dynamics

Jonathan E. Bronson^{*1}, Jake M. Hofman², Jingyi Fei¹, Ruben L. Gonzalez, Jr.¹ and Chris H. Wiggins³

¹Department of Chemistry, Columbia University, New York, NY 10027, USA

²Yahoo! Research, 111 West 40th St., New York, NY 10018, USA

³Department of Applied Physics and Applied Mathematics, Columbia University, New York, NY 10027, USA

Email: Jonathan E. Bronson^{*} - jeb2126@columbia.edu; Jake M. Hofman - jmh2045@columbia.edu; Jingyi Fei - jf2276@columbia.edu; Ruben L. Gonzalez, Jr. - rlg2118@columbia.edu; Chris H. Wiggins - chris.wiggins@columbia.edu;

^{*}Corresponding author

Abstract

Background: The recent explosion of experimental techniques in single molecule biophysics has generated a variety of novel time series data requiring equally novel computational tools for analysis and inference. This article describes in general terms how graphical modeling may be used to learn from biophysical time series data using the variational Bayesian expectation maximization algorithm (VBEM). The discussion is illustrated by the example of single-molecule fluorescence resonance energy transfer (smFRET) *versus* time data, where the smFRET time series is modeled as a hidden Markov model (HMM) with Gaussian observables. A detailed description of smFRET is provided as well.

Results: The VBEM algorithm returns the model's evidence and an approximating posterior parameter distribution given the data. The former provides a metric for model selection via maximum evidence (ME), and the latter a description of the model's parameters learned from the data. ME/VBEM provide several advantages over the more commonly used approach of maximum likelihood (ML) optimized by the expectation maximization (EM) algorithm, the most important being a natural form of model selection and a well-posed (non-divergent) optimization problem.

Conclusions: The results demonstrate the utility of graphical modeling for inference of dynamic processes in single molecule biophysics.

Background

Single-molecule techniques allow biophysicists to probe the dynamics of proteins, nucleic acids, and other biological macromolecules with unprecedented resolution [1–3]. It is now possible to observe viruses pack DNA into capsids [4], helicases unzip nucleic acids [5], motor proteins walk on biopolymers [6], and ribosome domains undergo structural rearrangements during translation [7]. These data are acquired by recording the fluorescent output or forces generated from, for example, biomolecules tethered onto microscope slides [8]; walking on biopolymers [9]; diffusing in hydrodynamic flow cells [10]; or pulled by optical [11] or magnetic [12] tweezers. Often the molecules studied move through a series of locally stable molecular conformations or positions (generically termed states) and give rise to data of the type shown in Fig. 1. From these data, the experimentalist wishes to learn a model describing the number of states occupied by the molecule and the transition rates between states. Although the myriad experimental techniques available have much in common, the data they generate often differ enough to require unique models.

For example, some of these models will involve conversion of chemical to mechanical energy, or motion associated with diffusion, or motion associated with transitions between distinct configurational states. Modeling the data, then, typically involves introducing several variables — some of which are observed, others of which are latent or “hidden”; some of which are real-valued coordinates, others of which are discrete states — and specifying algebraically how they are related. Such algebraic relations among a few variables are typical in physical modeling (e.g., the stochastic motion of a random walker, or the assumption of additive, independent, normally distributed errors typical in regression); models involving multiple conditionally-dependent observations or hidden variables with more structured noise behavior are less common. Implicitly, each equation of motion or of constraint specifies which variables are conditionally dependent and which are conditionally independent. Graphical modeling, which begins with charting these dependencies among a set of nodes, with edges corresponding to the conditional probabilities which must

be algebraically specified (*i.e.*, the typical elements of a physical model) organizes this process and facilitates basing inference on such models [13–15].

Here we explore the application of a specific subset of GMs to biophysical time series data using a specific algorithmic approach for inference: the directed GM and the variational Bayesian expectation maximization algorithm (VBEM). After discussing the theoretical basis and practical advantages of this general approach to modeling biophysical time series data, we apply the method to the problem of inference given single molecule fluorescence resonance energy transfer (smFRET) time series data. We emphasize the process and caveats of modeling smFRET data with a GM and demonstrate the most helpful features of VBEM for this type of time series inference.

Graphical models

GMs are a flexible inference framework based on factorizing a (high-dimensional) multivariate joint distribution into (lower-dimensional) conditionals and marginals [13–15]. In a GM, the nodes of the graph represent either observable variables (data, denoted by filled circles), latent variables (hidden states, denoted by open circles), or fixed parameters (denoted by dots). Directed edges between nodes represent conditional probabilities. For example, the three-node graphical model $X \rightarrow Y \rightarrow Z$ implies that the joint distribution $p(Z, Y, X) \equiv p(Z|Y, X)p(Y|X)p(X)$ can be further factorized as $p(Z|Y)p(Y|X)p(X)$. Data with a temporal component are modeled by connecting arrows from variables at earlier time steps to variables at later time steps. In many graphical models, the number of observed and latent variables grows with the size of the data set under consideration. To avoid clutter, these variables are written once and placed in a box, often called a “plate”, labeled with the number of times the variables are repeated [15]. This manuscript will denote hidden variables by z and observed data by d . Parameters which are vectors will be denoted as such by overhead arrows.

As an example of a simple GM, imagine trying to learn the number of boys and girls in an elementary school class of N students from a table of their heights and weights. Here the hidden variable is gender and the observed variable, (height, weight), is a random variable conditionally dependent on the hidden variable. The resulting GM is shown in Fig. 2, with the parameters of $p(\text{gender})$ denoted by $\vec{\alpha}$ and the

parameters $p(\text{height}, \text{weight} | \text{gender})$ denoted by $\vec{\mu}$ and $\vec{\Sigma}$. The expression for the probability of the observed data ($\{d_1, \dots, d_N\} = \mathbf{D}$) and latent genders ($\{z_1, \dots, z_N\} = \mathbf{Z}$) is uniquely specified by the graph and the factorization it implies:

$$p(\mathbf{D}, \mathbf{Z} | \vec{\mu}, \vec{\Sigma}, \vec{\alpha}) = \prod_{n=1}^N p(d_n | z_n, \vec{\mu}, \vec{\Sigma}) p(z_n | \vec{\alpha}). \quad (1)$$

In such a simple case it is straightforward to arrive at the expression in Eq. 1 without the use of a GM, but such a chart makes this factorization far more obvious and interpretable.

Inference of GMs

In some contexts, one wishes to learn the probability of the hidden states given the observed data, $p(\mathbf{Z} | \mathbf{D}, \vec{\theta}, K)$, where $\vec{\theta}$ denotes the parameters of the model and K denotes the number of allowed values of the latent variables (i.e. number of hidden states). If $\vec{\theta}$ is known then efficient inference of $p(\mathbf{Z} | \mathbf{D}, \vec{\theta}, K)$ can be performed on any loop-free graph with discrete latent states using the *sum-product* algorithm [16], or, if only the most probable values of \mathbf{Z} are needed, using the closely related *max-sum* algorithm [17]. A loop in a graph occurs when multiple pathways connect two variables, which is unlikely in a graph modeling time series data. Inference for models with continuous latent variables is discussed in [18, 19]. For most time series inference problems in biophysics, both \mathbf{Z} and $\vec{\theta}$ are unknown. In these cases, a criterion for choosing a best estimate of $\vec{\theta}$ and an optimization algorithm to find this estimate are needed.

Inference via maximum likelihood

Estimating $\vec{\theta}$ is most commonly accomplished using the *maximum likelihood* (ML) method, which estimates $\vec{\theta}$ as

$$\hat{\theta}_{\text{ML}} = \underset{\vec{\theta}}{\operatorname{argmax}} p(\mathbf{D} | \vec{\theta}, K) = \underset{\vec{\theta}}{\operatorname{argmax}} \sum_{\mathbf{Z}} p(\mathbf{D}, \mathbf{Z} | \vec{\theta}, K). \quad (2)$$

The probability $p(\mathbf{D} | \vec{\theta}, K)$ is known as the *likelihood*. The expectation maximization (EM) algorithm can be used to estimate $\hat{\theta}_{\text{ML}}$ [20]. In EM, an initial guess for $\hat{\theta}_{\text{ML}}$ is used to calculate $p(\mathbf{Z} | \mathbf{D}, \vec{\theta}, K)$. The $p(\mathbf{Z} | \mathbf{D}, \vec{\theta}, K)$ learned is then used to calculate a new guess for $\hat{\theta}_{\text{ML}}$. The algorithm iterates until convergence, and is guaranteed to converge to a local optimum. The EM algorithm should be run with

multiple initializations of $\hat{\theta}_{\text{ML}}$, often called “random restarts”, to increase the probability of finding the globally optimal $\hat{\theta}_{\text{ML}}$.

ML solved via EM is a generally effective method to perform inference however, it has two prominent shortcomings [14, 15]:

Model selection: The first limitation of ML is that it has no form of model selection: the likelihood monotonically increases with the addition of more model parameters. This problem of fitting too many states to the data (overfitting) is highly undesirable for biophysical time series data, where learning the correct K for the data is often an experimental objective.

Ill-posedness The second problem with ML occurs only in the case of a model with multiple hidden states and a continuous observable (a case which includes the majority of biophysical time series data, including the smFRET data we will consider here). If the mean of one hidden state approaches the position of a data point and the variance of that state approaches zero, the contribution of that datum to the likelihood will diverge. When this happens, the likelihood will be infinite regardless of how poorly the rest of the data are modeled, provided the other states in the model have non-zero probabilities for the rest of the data. For such models, the ML method is ill-posed; poor parameters can still result in infinite likelihood.

In practical contexts, the second problem (divergent likelihood) can be avoided either by performing MAP estimation (maximizing the likelihood times a prior which penalizes small variance) or by ignoring solutions for which likelihood is diverging. That is, one does not actually maximize the likelihood. Model selection can then be argued for based on cross-validation or by penalizing likelihood with a term which monotonically increases with model complexity [15, 21, 22]. We consider, instead, an alternative optimization criterion which naturally avoids these problems.

Inference via maximum evidence

A Bayesian alternative to ML is to perform inference using the *maximum evidence* (ME) method. ME can be thought of as an extension of ML to the problem of model selection. Where ML asks which parameters

maximize the probability of the data for a given model, ME asks which model, including nested models which differ only in K , makes the data most probable. According to ME, the model of correct complexity (K_*) is

$$K_* = \operatorname{argmax}_K p(\mathbf{D}|\vec{u}, K) = \operatorname{argmax}_K \sum_{\mathbf{Z}} \int d\vec{\theta} p(\mathbf{D}, \mathbf{Z}|\vec{\theta}, K) p(\vec{\theta}|\vec{u}, K). \quad (3)$$

The quantity $p(\mathbf{D}|\vec{u}, K)$ is called the evidence. Sometimes it is also referred to as the marginal likelihood, since unknown parameters are assigned probability distributions and marginalized (or summed out) over all possible settings. The evidence penalizes both models which underfit and models which overfit. The second expression in Eq. 3 follows readily from the sum rule of probability provided we are willing to model the parameters themselves as random variables. That is, we are willing to specify a distribution over parameters, $p(\vec{\theta}|\vec{u}, K)$. This distribution is called the “prior”, since it can be thought of as the probability of the parameters prior to seeing any data. The parameters for the distributions of the prior (\vec{u}) are called *hyperparameters*. In addition to providing a method for model selection, by integrating over parameters to calculate the evidence rather than using a “best” point estimate of the parameters, ME avoids the ill-posedness problem associated with ML.

Although ME provides an approach to model selection, calculation of the evidence does not, on its own, provide an estimate for $\vec{\theta}$. The VBEM approach to estimating evidence does, however, provide a mechanism to estimate $\vec{\theta}$. VBEM can be thought of as an extension of EM for ME. Both the VBEM algorithm and considerations for choosing priors are discussed in Methods.

smFRET

Before building a GM describing smFRET data, it is helpful to review briefly the experimental method. The experimental technique is based on the spectroscopic phenomenon that, if the emission spectrum of a polar chromophore (donor) overlaps with the absorption spectrum of another polar chromophore (acceptor), electromagnetic excitation of the donor can induce a transfer of energy to the acceptor via a non-radiative, dipole-dipole coupling process termed fluorescence resonance energy transfer (FRET) [23]. The transfer efficiency between donor and acceptor scales with the distance between molecules (r) as $1/r^6$, with FRET efficiencies most sensitive to r in the range of 1 – 10nm. Because of this extraordinary sensitivity to distance, FRET efficiency can serve as a molecular ruler, allowing an experimentalist to

measure the separation between donor and acceptor by stimulating the donor with light and measuring emission intensities of both the donor (I_D) and acceptor (I_A) [24]. Usually a summary statistic called the “FRET ratio” is used to report on molecular distance rather than the “raw”, 2-channel $\{I_A, I_D\}$ data, although inference of the raw 2-channel data is possible as well [25]. The FRET ratio is given by

$$FRET = \frac{I_A}{I_D + I_A}. \quad (4)$$

When the donor and acceptor are attached to an individual protein, nucleic acid, or other molecular complex, the FRET signal can be used to report on the dynamics of the molecule to which the donor and acceptor are attached (see Fig. 3). When the experiment is crafted to monitor individual molecules rather than ensembles of molecules, the process is termed single molecule FRET (smFRET). For many biological studies, such as the identification and characterization of the structural dynamics of a biomolecule, smFRET must be used rather than FRET; the majority of molecular dynamics cannot be observed from ensemble averages. Often the molecule of interest adopts a series of locally stable conformations during a smFRET time series. From these data, the experimentalist would like to learn (1) the number of locally stable conformations in the data (i.e. states) and (2) the transition rates between states. Although it is theoretically possible use the FRET signal to quantify the distance between parts of a molecule during a time series, there are usually too many variables affecting FRET efficiency for this to be practical [26]. Consequently, smFRET is usually used to extract quantitative information about kinetics (i.e. rate constants) but only qualitative information about distances.

The photophysics of FRET have been studied for over half a century, but the first smFRET experiments were only carried out about fifteen years ago [27]. The field has been growing exponentially since, and hundreds of smFRET papers are published annually [1]. Diverse topics such as protein folding [28], RNA structural dynamics [29], and DNA-protein interactions [30] have been investigated via smFRET. The size and complexity of smFRET experiments has grown substantially since the original smFRET publication. A modern smFRET experiment can generate thousands of time series to be analyzed [7].

Results and Discussion

smFRET as a graphical model

A model of the smFRET time series for a molecule transitioning between a series of locally stable conformations should capture several important aspects of the process [31]. The observable smFRET signal is a function of the hidden conformation of the molecule. The noise of each smFRET state can be assumed to be Gaussian, and the hidden conformations are assumed to be discrete and finite in number. The probability of transitioning to a new molecular conformation should be a function of the current conformation of the molecule (*e.g.*, the DNA in Fig. 3 is more likely to be zipped at time $t + 1$ if it is zipped at time t). The CCD cameras commonly used in smFRET experiments naturally bin the data temporally, so it is convenient to work with a model where time is discrete. The GM expressing these features is called a hidden Markov model (HMM) and is shown in Fig. 4A. From the graph, it can be seen that the probability of the observed and latent variables factorizes as

$$p(\mathbf{D}, \mathbf{Z} | \vec{\theta}, K) = p(z_1 | \vec{\theta}, K) \left[\prod_{t=2}^T p(z_t | z_{t-1}, \vec{\theta}, K) \right] \prod_{t=1}^T p(d_t | z_t, \vec{\theta}, K). \quad (5)$$

Here, $\vec{\theta}$ must include parameters for the probability that the time series begins in each state ($p(z_1 = k) \equiv \pi_k$); parameters for transition probabilities between states ($p(z_{t+1} = k | z_t = j) \equiv a_{jk}$); and parameters for the noise of the emissions of each state ($p(d_t | z_t = k) = \mathcal{N}(d_t | \mu_k, \lambda_k)$, where μ_k and λ_k are the mean and precision of the Gaussian). It is necessary to model $p(z_1)$ separately from all other transition probabilities since it is the only hidden state probability which does not depend on z_{t-1} . The a_{jk} are commonly represented as a matrix, A , called a transition matrix. The probability the time series begins in the k^{th} state and transition probabilities between states are drawn from multinomial distributions defined by $\vec{\pi}$ and the rows of A , respectively. The GM for this HMM is shown in Fig. 4B. From the GM it can be seen that

$$\begin{aligned} p(\mathbf{D}, \mathbf{Z} | \vec{\theta}, K) p(\vec{\theta} | \vec{u}, K) &= p(z_1 | \vec{\pi}, K) \left[\prod_{t=2}^T p(z_t | z_{t-1}, A, K) \right] \prod_{t=1}^T p(d_t | z_t, \vec{\mu}, \vec{\lambda}, K) \times \\ &\quad p(\vec{\pi} | u_\pi, K) p(A | \vec{u}_A, K) p(\vec{\mu} | \vec{u}_m, \vec{u}_\beta, \vec{\lambda}, K) p(\vec{\lambda} | \vec{u}_a, \vec{u}_b, K). \end{aligned} \quad (6)$$

For a time series of length T where each latent variable can take on K states, a brute summation over all

possible states requires $O(K^T)$ calculations. By exploiting efficiencies in the GM and using the sum-product algorithm, this summation can be performed using $O(K^2T)$ calculations (which can be seen by noting that the latent state probabilities in Eq. 6 factorize into $p(z_t|z_{t-1}, A, K)$, where each of the T latent states has K^2 possible combinations of states). The sum-product algorithm applied to the HMM is called the forward-backward algorithm or the Baum-Welch algorithm [32], and the most probable trajectory is called the Viterbi path [33].

There are several assumptions of this model which should be considered. First, although it is common to assume the noise of smFRET states is Gaussian, the assumption does not have a theoretical justification (and since FRET intensities can only be on the interval $(0, 1)$, and the Gaussian distribution has support $(-\infty, \infty)$, the data cannot be truly Gaussian). Despite this caveat, several groups have successfully modeled smFRET the data as having Gaussian states [25, 34, 35]. We note that other distributions have been considered as well [36].

Second, the HMM assumes that the molecule instantly switches between hidden states. If the time it takes the molecule to transition between conformations is on the same (or similar) order of magnitude as the time it spends within a conformation, the HMM is not an appropriate model for the process and a different GM will be needed. For many molecular processes, such as protein domain rearrangements, the molecule transitions between conformations orders of magnitude faster than it remains in a conformation and the HMM can model the process well [37].

Third, the HMM is “memoryless” in the sense that, given its current state, the transition probabilities are independent of the past. It is still possible to model a molecule which sometimes transitions between states quickly and sometimes transitions between states slowly (if, for example, binding of another small molecule to the molecule being studied changes its transition rates [7]). This situation can be modeled using two latent states for each smFRET state. The two latent states will have the same emissions model parameters, but different transition rates.

Illustration of the inference

A software package, vbFRET, implementing the VBEM algorithm for this HMM was written and described in [25], along with an assessment of the algorithm’s performance on real and synthetic data. An illustration of the method is shown here, demonstrating three of its most important abilities: the ability to perform model selection; the ability to learn posterior parameter distributions; and the ability to idealize a time series. These abilities are demonstrated on three synthetic $K = 3$ state time series, shown in Fig. 5C. The traces all have $\mu = \{0.25, 0.5, 0.75\}$ and identical hidden state trajectories. The noise of each hidden state is $\sigma = 0.015$ for trace 1 (unrealistically noiseless), $\sigma = 0.09$ for trace 2 (a level of noise commonly encountered in experiments), and $\sigma = 0.15$ for trace 3 (unrealistically noisy).

Model selection: For each trace, $\mathcal{L}(q)$, the lower bound of the log(evidence), was calculated for $1 \leq K \leq 7$. The results are shown in Fig. 5A, with the largest value of $\mathcal{L}(q)$ for each trace shown in red. For traces 1 and 2, $\mathcal{L}(q)$ peaks for $K_* = 3$, correctly inferring the complexity of the model. For trace 3, the noise of the system is too large, given the length of the trace, to infer three clearly resolved states. For this trace $\mathcal{L}(q)$ peaks at $K_* = 2$. This result illustrates an important consideration of evidence based model selection: states which are distinct in a generative model (or an experiment generating data) may not give rise to statistically significant states in the data generated. For example, two states which have identical means, variances, and transition rates would be statistically indistinguishable from a single state with those parameters. When states are resolvable, however, ME-based model selection is generally effective, as demonstrated in traces 1 and 2.

Posterior distributions: The ability to learn a complete posterior distribution for $\vec{\theta}$ provides more information than simply learning an estimate for $\vec{\theta}$, and is a feature unique to Bayesian statistics. The maximum of the distribution, denoted $\hat{\theta}_{\text{MAP}}$, can be used as an estimate of $\vec{\theta}$ (e.g., if idealized trajectories are needed). The subscript here differentiates it from the estimate in the absence of the prior, $\hat{\theta}_{\text{ML}}$. The curvature of the distribution describes the certainty of the $\hat{\theta}_{\text{MAP}}$ estimate. As a demonstration, the posterior for the mean of the lowest smFRET state of each trace is shown in Fig. 5B. The X and Y axes are the same in all three plots, so the distributions can be compared. As expected, the lower the noise in the trace, the narrower the posterior distribution and the higher the confidence of the estimate for μ . The estimate of μ for trace 3 is larger than in the other traces because $K_* = 2$; some the middle smFRET state

data are misclassified as belonging to the low smFRET state.

Idealized trajectories: Idealized smFRET trajectories can be a useful visual aid to report on inference. They are also a necessity for some forms of post-processing commonly used at present, such as dwell-time analysis [7]. Idealized trajectories can be generated from the posterior learned from VBEM by using $\hat{\theta}_{\text{MAP}}$ to calculate the most probable hidden state trajectory (the Viterbi path) [33]. The idealized trajectories for each trace are shown in Fig. 5C. For traces 1 and 2, where K_* is correctly identified, the idealized trajectory captures the true hidden state trajectory perfectly. Because of the model selection and well-posedness of ME/VBEM the idealized trajectories learned with this method can be substantially more accurate than those learned by ML for some data sets [25].

Conclusions

This manuscript demonstrates how graphical modeling, in conjunction with a detailed description of a biophysical process, can be used to model biophysical time series data effectively. The GM designed here is able to model smFRET data and learn both the number of states in the data and the posterior parameter values for those states. The ME/VBEM methodology used here offers several advantages over the more commonly used ML/EM inference approach, including intrinsic model selection and a well-posed optimization. All modeling assumptions are readily apparent from the GM. The GM framework with inference using ME/VBEM is highly flexible modeling approach which we anticipate will be applicable to a wide array of problems in biophysics.

Methods

All code used in this manuscript is available open source at <http://vbfret.sourceforge.net/>.

Variational Bayesian expectation maximization

Unfortunately, calculation of Eq. 3 requires a sum over all K settings for each of T extensive variables \mathbf{Z} (where T is the length of the time series). Such a calculation is numerically intractable, even for reasonably small systems (*e.g.*, $K=2$, $T=100$) so an approximation to the evidence must be used. Several approximation methods exist, such as Monte Carlo techniques, for numerically approximating such sums [38]. The method we will consider here is VBEM.

One motivation for the VBEM algorithm is the following simple algebraic identity [15]. Since Bayesian analysis treats latent variables (\mathbf{Z}) and unknown parameters ($\vec{\theta}$) the same way this section will lump them both into \mathbf{X} for notational simplicity. Let $q(\mathbf{X})$ be any probability distribution over \mathbf{X} . Then,

$$\log p(\mathbf{D}|\vec{u}, K) = \int q(\mathbf{X}) \log (p(\mathbf{D}|\vec{u}, K)) d\mathbf{X} \quad (7)$$

$$= \int q(\mathbf{X}) \log \left(\frac{p(\mathbf{D}, \mathbf{X}|\vec{u}, K)}{p(\mathbf{X}|\mathbf{D}, \vec{u}, K)} \right) d\mathbf{X} \quad (8)$$

$$= \int q(\mathbf{X}) \log \left(\frac{p(\mathbf{D}, \mathbf{X}|\vec{u}, K)q(\mathbf{X})}{p(\mathbf{Z}|\mathbf{D}, \vec{u}, K)q(\mathbf{X})} \right) d\mathbf{X} \quad (9)$$

$$= \int q(\mathbf{X}) \log \left(\frac{p(\mathbf{D}, \mathbf{X}|\vec{u}, K)}{q(\mathbf{X})} \right) d\mathbf{X} \\ - \int q(\mathbf{X}) \log \left(\frac{p(\mathbf{X}|\mathbf{D}, \vec{u}, K)}{q(\mathbf{X})} \right) d\mathbf{X} \quad (10)$$

$$= \mathcal{L}(q) + D_{KL} \left(q(\mathbf{Z}, \vec{\theta}) || p(\mathbf{Z}, \vec{\theta}|\mathbf{D}, \vec{u}, K) \right). \quad (11)$$

Summations over the discrete components of \mathbf{X} should be included in these equations, but are omitted for notational simplicity. The equality in Eq. 7 results from the requirement that $q(\mathbf{X})$ be a normalized probability; Eq. 8 rewrites $p(\mathbf{D}|\vec{u}, K)$ in terms of a conditional probability; and Eq. 11 reinserts $\{\mathbf{Z}, \vec{\theta}\}$ for \mathbf{X} and renames the two terms in Eq. 10 as $\mathcal{L}(q)$, the lower bound of the log(evidence), and the Kullback–Leibler divergence, respectively.

Using Jensen’s inequality, it can be shown that

$$D_{KL} (q||p) \geq 0, \quad (12)$$

with equality when $q = p$. Consequently,

$$\log (p(\mathbf{D}|\vec{u}, K)) \geq \mathcal{L}(q), \quad (13)$$

i.e., $\exp(\mathcal{L}(q))$ is a lower bound on the model's evidence. Eq. 12 implies that $\mathcal{L}(q)$ is maximized when $q(\mathbf{Z}, \vec{\theta})$ is equal to $p(\mathbf{Z}, \vec{\theta} | \mathbf{D}, \vec{u}, K)$. As a corollary, from this it follows that $q(\theta)$ approximates $p(\vec{\theta} | \mathbf{D}, \vec{u}, K)$, the *posterior* distribution of the parameters. Therefore, the optimization simultaneously performs model selection (by finding a K which maximizes $p(\mathbf{D} | \vec{u}, K)$) and inference (by approximating $p(\mathbf{Z}, \vec{\theta} | \mathbf{D}, \vec{u}, K)$).

The approach suggested by Eqs. 7–12 is to replace an intractable calculation with a tractable bound optimization. If $p(\mathbf{D} | \vec{\theta}, K)$ is in the exponential family and a conjugate prior is used, then the only assumption about $q(\mathbf{Z}, \vec{\theta})$ needed is that $q(\mathbf{Z}, \vec{\theta}) = q(\mathbf{Z})q(\vec{\theta})$ (*i.e.*, it factorizes into a function of \mathbf{Z} and a function of $\vec{\theta}$) for the inference problem to be tractable using VBEM [39]. In addition, under these conditions $p(\vec{\theta} | \mathbf{D}, \vec{u}, K)$ will have the same functional form as $p(\vec{\theta} | \vec{u}, K)$. The VBEM algorithm is similar to EM, but rather than iteratively using guesses for $\hat{\theta}_{\text{ML}}$ to set \mathbf{Z} and guesses for \mathbf{Z} to set $\hat{\theta}_{\text{ML}}$ the update equations iterate between [15, 40]:

$$\text{VBE} : q(\mathbf{Z}) = \frac{1}{\mathcal{Z}_{\mathbf{Z}}} \exp \left(\mathbb{E}_{q(\vec{\theta})} \left[\log \left(p(\mathbf{D}, \mathbf{Z} | \vec{\theta}, K) p(\vec{\theta} | \vec{u}, K) \right) \right] \right) \quad (14)$$

$$\text{VBM} : q(\vec{\theta}) = \frac{1}{\mathcal{Z}_{\vec{\theta}}} \exp \left(\mathbb{E}_{q(\mathbf{Z})} \left[\log \left(p(\mathbf{D}, \mathbf{Z} | \vec{\theta}, K) p(\vec{\theta} | \vec{u}, K) \right) \right] \right). \quad (15)$$

Here \mathbb{E} denotes the expected value with respect to the subscripted distribution and \mathcal{Z} is a normalization constant. Whereas the $\log(p(\mathbf{D} | \vec{u}, K))$ is a log of a sum/integral, the right hand sides of Eqs. 14 & 15 both involve the sum/integral of a log. This difference renders $\log(p(\mathbf{D} | \vec{u}, K))$ intractable, yet Eqs. 14 & 15 tractable.

An interesting and potentially useful feature of the $q(\vec{\theta})$ learned from VBEM is that when K is chosen to be larger than the number of states supported by the data, the optimization leaves the extra states unpopulated. This propensity to leave unnecessary states unpopulated in the posterior, sometimes called “extinguishing”, is a second form of model selection intrinsic to VBEM, which is in addition to the model selection described by Eq. 3. An explanation for this behavior can be found in Chapter 3 of [15].

Priors

Several considerations should go into choosing a prior. Choosing distributions which are conjugate to the parameters of the likelihood can greatly simplify inference [39]. Priors can be chosen to minimize their

influence on the inference. Such priors are called “weak” or uninformative. Alternatively, priors can also be chosen to respect previously obtained experimental observations [40]. It is important to check that inference results do not heavily depend on the prior (*e.g.* doubling or halving hyperparameter values should not affect inference results).

The conjugate prior of a multinomial distribution is a Dirichlet distribution: $p(\pi_1, \dots, \pi_K) = \text{Dir}(\vec{\pi}|u_\pi)$; $p(a_{k1}, \dots, a_{kK}) = \text{Dir}(a_{k,1-K}|u_A^k)$. Expressed in terms of precision λ , rather than variance σ^2 (where $\lambda = 1/\sigma^2$), the conjugate prior for the mean and precision of a Gaussian is a Gaussian-Gamma distribution: $p(\mu_k, \lambda_k) = \mathcal{N}(\mu_k|u_m^k, (u_\beta^k \lambda_k)^{-1})\text{Gam}(\lambda_k|u_a^k, u_b^k)$.

Here, hyperparameters were set so as to give distributions consistent with experimental data and to influence the inference as weakly as possible: $u_\pi^k = 1$, $u_a^{jk} = 1$, $u_\beta^k = 0.25$, $u_m^k = 0.5$, $u_a^k = 2.5$ and $u_b^k = 0.01$, for all values of k . Qualitatively, these hyperparameters specify probability distributions over the hidden states such that it is most probable that the hidden states are equally likely to be occupied and equally likely to be transitioned to. Quantitatively, they yield $\langle \mu_k \rangle = 0.5$ and $\text{mode}[\sigma] \approx 0.08$, consistent with experimental observation:

$$\frac{1}{\sqrt{\text{mode}[\lambda_k]}} = \sqrt{\frac{u_b^k}{(u_a^k - 1)}} \approx 0.08 \quad \forall k. \quad (16)$$

Data generation

Synthetic traces were generated in MATLAB using 1-D Gaussian noise for each hidden state and a manually determined hidden state trajectory. All traces were analyzed by vbFRET [25], using its default parameter settings, for $1 \geq K \geq 7$, with 25 random restarts for each value of K . The restart with the highest evidence was used to generate the data in Fig. 5. The posterior probability of μ_k is given by $\mathcal{N}(\mu_k|v_m^k, (v_\beta^k \lambda_k)^{-1})$, where \vec{v} are the hyperparameters of the posterior. The data in Fig. 5B were generated using this equation with λ_k fixed at its most probable posterior value.

Competing interests

The authors declare that they have no competing interests.

Authors contributions

JEB contributed to the graphical modeling and smFRET inference. JMH contributed to the graphical modeling. JF and RLG contributed to the smFRET inference. CHW contributed to the graphical modeling and smFRET inference. JEB and CHW wrote the manuscript.

Acknowledgments

This work was supported by a grant to CHW from the NIH (5PN2EY016586-03); grants to RLG from the Burroughs Wellcome Fund (CABS 1004856), the NSF (MCB 0644262), and the NIH-NIGMS (1RO1GM084288-01); and a grant to JEB from the NSF (GRFP).

References

1. Joo C, Balci H, Ishitsuka Y, Buranachai C, Ha T: **Advances in single-molecule fluorescence methods for molecular biology**. *Annu. Rev. Biochem.* 2008, **77**:51–76.
2. Myong S, Ha T: **Stepwise translocation of nucleic acid motors**. *Curr. Opin. Struct. Biol.* 2010, **20**:121–127.
3. Seidel R, Dekker C: **Single-molecule studies of nucleic acid motors**. *Curr. Opin. Struct. Biol.* 2007, **17**:80–86.
4. Aathavan K, Politzer AT, Kaplan A, Moffitt JR, Chemla YR, Grimes S, Jardine PJ, Anderson DL, Bustamante C: **Substrate interactions and promiscuity in a viral DNA packaging motor**. *Nature* 2009, **461**:669–673.
5. Dumont S, Cheng W, Serebrov V, Beran RK, Tinoco I, Pyle AM, Bustamante C: **RNA translocation and unwinding mechanism of HCV NS3 helicase and its coordination by ATP**. *Nature* 2006, **439**:105–108.
6. Mori T, Vale RD, Tomishige M: **How kinesin waits between steps**. *Nature* 2007, **450**:750–754.
7. Fei J, Bronson JE, Hofman JM, Srinivas RL, Wiggins CH, Gonzalez RL: **Allosteric collaboration between elongation factor G and the ribosomal L1 stalk directs tRNA movements during translation**. *Proc. Natl. Acad. Sci. U.S.A.* 2009, **106**:15702–15707.
8. Fei J, Kosuri P, MacDougall DD, Gonzalez RL: **Coupling of ribosomal L1 stalk and tRNA dynamics during translation elongation**. *Mol. Cell* 2008, **30**:348–359.

9. Block SM, Goldstein LS, Schnapp BJ: **Bead movement by single kinesin molecules studied with optical tweezers.** *Nature* 1990, **348**:348–352.
10. Visnapuu ML, Greene EC: **Single-molecule imaging of DNA curtains reveals intrinsic energy landscapes for nucleosome deposition.** *Nat. Struct. Mol. Biol.* 2009, **16**:1056–1062.
11. Perkins TT, Quake SR, Smith DE, Chu S: **Relaxation of a single DNA molecule observed by optical microscopy.** *Science* 1994, **264**:822–826.
12. Yan J, Skoko D, Marko JF: **Near-field-magnetic-tweezer manipulation of single DNA molecules.** *Phys Rev E Stat Nonlin Soft Matter Phys* 2004, **70**:011905.
13. Jordan M, Ghahramani Z, Jaakkola T, Saul L: **An introduction to variational methods for graphical models.** *Machine Learning* 1999, **37**(2):183–233.
14. MacKay DJ: *Information theory, inference, and learning algorithms.* Cambridge University Press 2003.
15. Bishop C: *Pattern Recognition and Machine Learning.* Oxford Oxfordshire: Oxford University Press 2006.
16. Kschischang F, Frey B, Loeliger H: **Factor graphs and the sum-product algorithm.** *IEEE Transactions on Information Theory* 2001, **47**(2):498–519.
17. Weiss Y, Freeman W: **On the optimality of solutions of the max-product belief-propagation algorithm in arbitrary graphs.** *IEEE Transactions on Information Theory* 2001, **47**(2):736–744.
18. Ghahramani Z, Beal M: **Propagation Algorithms for Variational Bayesian Learning.** In *Advances in Neural Information Processing Systems 13*, Cambridge, MA: MIT Press 2001.
19. Bishop C, Spiegelhalter D, Winn J: **VIBES: A Variational Inference Engine for Bayesian Networks.** In *Advances in Neural Information Processing Systems 15*, Cambridge, MA: MIT Press 2003.
20. Dempster AP, Laird NM, Rubin DB: **Maximum likelihood from incomplete data via EM algorithm.** *Journal of the Royal Statistical Society Series B-Methodological* 1977, **39**(1):1–38.
21. Akaike H: **A new look at statistical-model identification.** *IEEE Transactions on Automatic Control* 1974, **AC19**(6):716–723.
22. Schwarz G: **Estimating the dimension of a model.** *The Annals of Statistics* 1978, **6**(2):461–464.
23. Förster T: **Zwischenmolekulare Energiewanderung Und Fluoreszenz.** *Annalen Der Physik* 1948, **2**(1-2):55–75.
24. Stryer L, Haugland RP: **Energy transfer: a spectroscopic ruler.** *Proc. Natl. Acad. Sci. U.S.A.* 1967, **58**:719–726.
25. Bronson JE, Fei J, Hofman JM, Gonzalez RL, Wiggins CH: **Learning rates and states from biophysical time series: a Bayesian approach to model selection and single-molecule FRET data.** *Biophys. J.* 2009, **97**:3196–3205.
26. Schuler B, Lipman EA, Steinbach PJ, Kumke M, Eaton WA: **Polyproline and the "spectroscopic ruler" revisited with single-molecule fluorescence.** *Proc. Natl. Acad. Sci. U.S.A.* 2005, **102**:2754–2759.
27. Ha T, Enderle T, Ogletree DF, Chemla DS, Selvin PR, Weiss S: **Probing the interaction between two single molecules: fluorescence resonance energy transfer between a single donor and a single acceptor.** *Proc. Natl. Acad. Sci. U.S.A.* 1996, **93**:6264–6268.
28. Deniz AA, Laurence TA, Beligere GS, Dahan M, Martin AB, Chemla DS, Dawson PE, Schultz PG, Weiss S: **Single-molecule protein folding: diffusion fluorescence resonance energy transfer studies of the denaturation of chymotrypsin inhibitor 2.** *Proc. Natl. Acad. Sci. U.S.A.* 2000, **97**:5179–5184.
29. Zhuang X, Kim H, Pereira MJ, Babcock HP, Walter NG, Chu S: **Correlating structural dynamics and function in single ribozyme molecules.** *Science* 2002, **296**:1473–1476.
30. Roy R, Kozlov AG, Lohman TM, Ha T: **SSB protein diffusion on single-stranded DNA stimulates RecA filament formation.** *Nature* 2009, **461**:1092–1097.
31. Andrec M, Levy RM, Talaga DS: **Direct Determination of Kinetic Rates from Single-Molecule Photon Arrival Trajectories Using Hidden Markov Models.** *J Phys Chem A* 2003, **107**:7454–7464.
32. Rabiner LR: **A Tutorial On Hidden Markov-Models And Selected Applications In Speech Recognition.** *Proceedings of the Ieee* 1989, **77**(2):257–286.

33. Viterbi AJ: **Error Bounds For Convolutional Codes And An Asymptotically Optimum Decoding Algorithm.** *IEEE Transactions On Information Theory* 1967, **13**(2):260+.
34. Qin F, Auerbach A, Sachs F: **Maximum likelihood estimation of aggregated Markov processes.** *Proceedings of the Royal Society of London Series B-Biological Sciences* 1997, **264**(1380):375–383.
35. McKinney SA, Joo C, Ha T: **Analysis of single-molecule FRET trajectories using hidden Markov modeling.** *Biophys. J.* 2006, **91**:1941–1951.
36. Liu Y, Park J, Dahmen KA, Chemla YR, Ha T: **A comparative study of multivariate and univariate hidden Markov modelings in time-binned single-molecule FRET data analysis.** *J Phys Chem B* 2010, **114**:5386–5403.
37. Creighton TE: *Proteins: Structures and Molecular Properties.* W. H. Freeman 1992.
38. Neal R: **Probabilistic inference using Markov chain Monte Carlo methods.** *Technical Report CRG-TR-93-1, Department of Computer Science, University of Toronto* 1993.
39. Beal MJ, Ghahramani Z: **Variational Bayesian Learning of Directed Graphical Models with Hidden Variables.** *Bayesian Analysis* 2006, **1**(4):793–831.
40. Beal M: **Variational Algorithms for Approximate Bayesian Inference.** *PhD thesis, University of Cambridge, UK*, <http://www.cse.buffalo.edu/faculty/mbeal/papers.html> 2003.

Figures

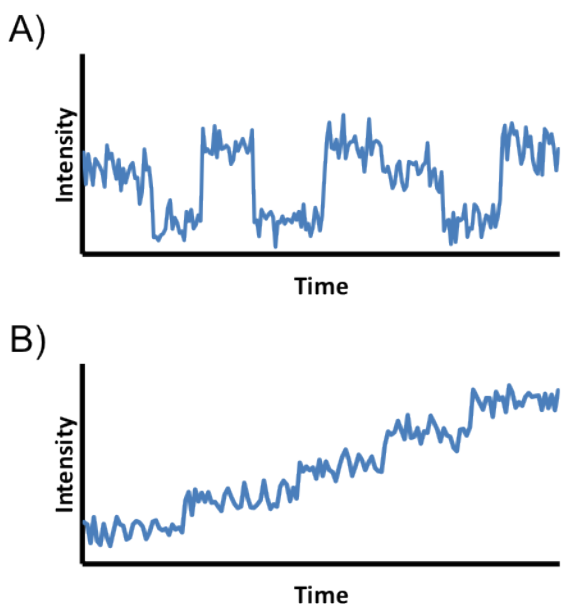


Figure 1: Examples of types of commonly encountered biophysical time series data. (A) A time series for a molecule transitioning between a series of locally stable conformations. Such data often arise, for example, when studying protein domain movements or the dynamics of polymers tethered to a surface. (B) A time series for a molecule undergoing a stepping process. Such data often arise, for example, when studying proteins with unidirectional movements, e.g., helicases and motor proteins.

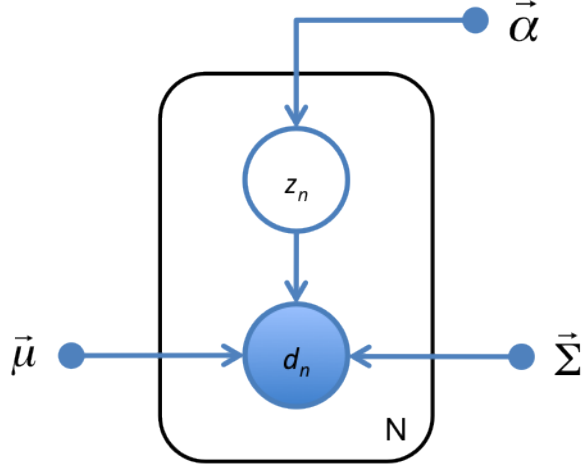


Figure 2: A GM for the problem of learning genders of boys and girls from a table of their heights and weights. The gender of the n^{th} child is denoted z_n . The 2-dimensional vector of the child's height and weight is denoted d_n . The mean height and weight for each gender, variances of height and weight for each gender, and probability of belonging to each gender are denoted by $\vec{\mu}$, $\vec{\Sigma}$, and $\vec{\alpha}$, respectively. Observed variables are represented by open circles, hidden variables are represented by filled circles, and fixed parameters are represented by dots. To avoid drawing nodes for all N hidden and observed variables, the variables are shown once and placed inside a plate which denotes the number of repetitions in the lower right corner. This GM specifies the conditional factorization of $p(\mathbf{D}, \mathbf{Z}, \vec{\theta})$ shown in Eq. 1.

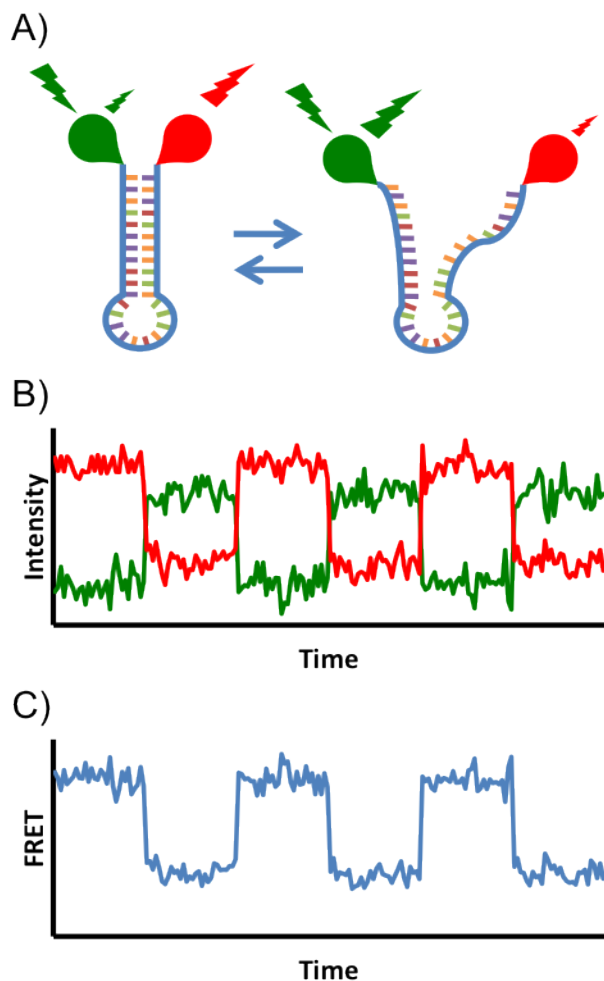


Figure 3: (A) Cartoon of a smFRET experiment studying the zipping/unzipping of a DNA hairpin. A FRET donor chromophore (green balloon) and acceptor chromophore (red balloon) are attached to the DNA. When the DNA is zipped (left), exciting the donor with green light causes the majority of energy to be transferred to the acceptor. The donor will fluoresce dimly and the acceptor will fluoresce brightly. When the DNA is unzipped, the probes are too far apart for efficient FRET. Exciting the donor in this conformation causes it to fluoresce brightly and the acceptor to fluoresce dimly. (B) The two channel (donor, acceptor) time series generated by the DNA as it transitions between zipped (bright red, dim green) and unzipped (dim red, bright green). (C) The 1D FRET transformation of the time series from B, calculated with Eq. 4. The closer the probes, the more intense the signal. Time series of this summary statistic are commonly used for analysis.

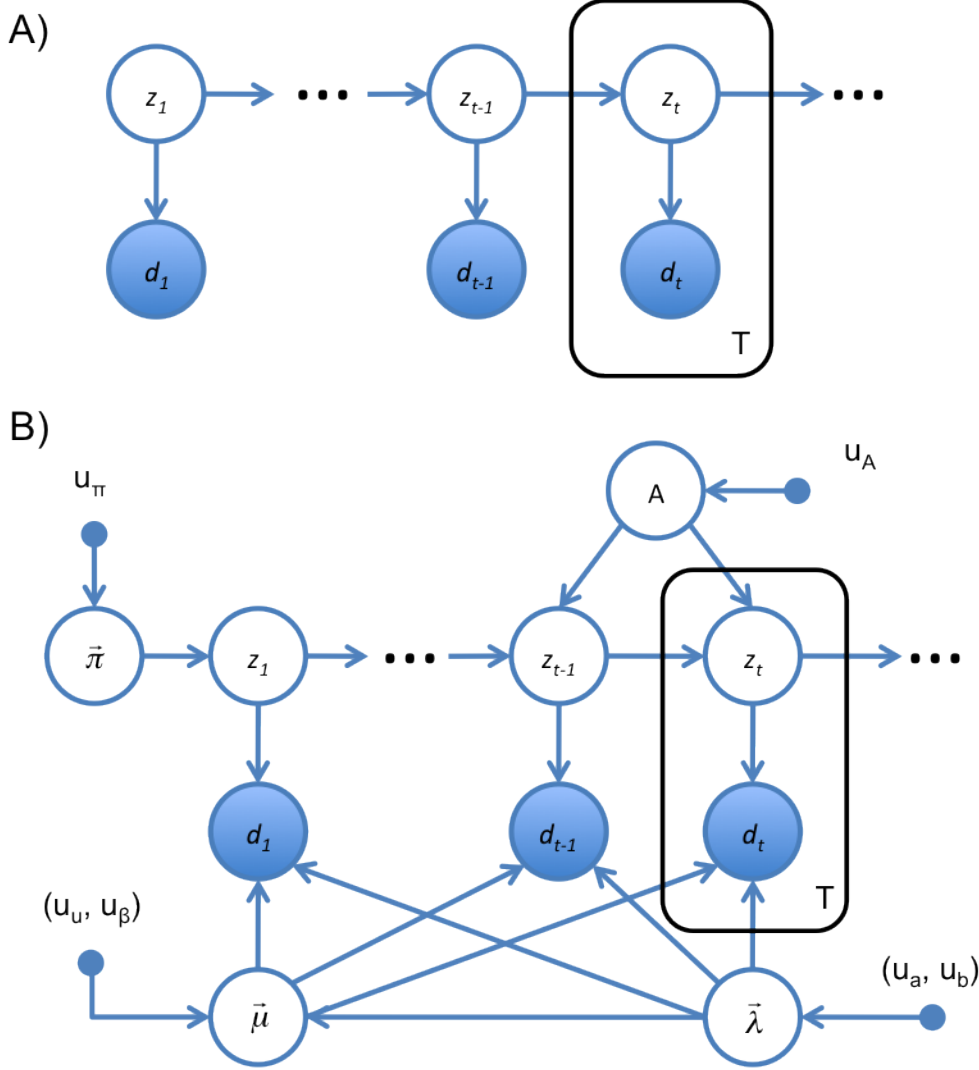


Figure 4: (A) The HMM as a GM. At each time step, t , the system occupies a hidden state, z_t and produces an observable emission, d_t , drawn from $p(d_t|z_t)$. In turn, z_t is drawn from $p(z_t|z_{t-1})$. (B) Complete GM for the HMM used to describe smFRET data in this work. Following the Bayesian treatment of probability, all unknown parameters are treated as hidden variables, and represented as open circles. Emissions are assumed to be Gaussian, with mean $\bar{\mu}$ and precision $\bar{\lambda}$. Transition rates are multinomial, with probabilities given by A . The probability of initially occupying each hidden state is multinomial as well, with probabilities given by $\bar{\pi}$. Equations for these distributions are described in the text below Eq. 5. This GM specifies the conditional factorization of $p(\mathbf{D}, \mathbf{Z}, \bar{\theta})$ shown in Eq. 6.

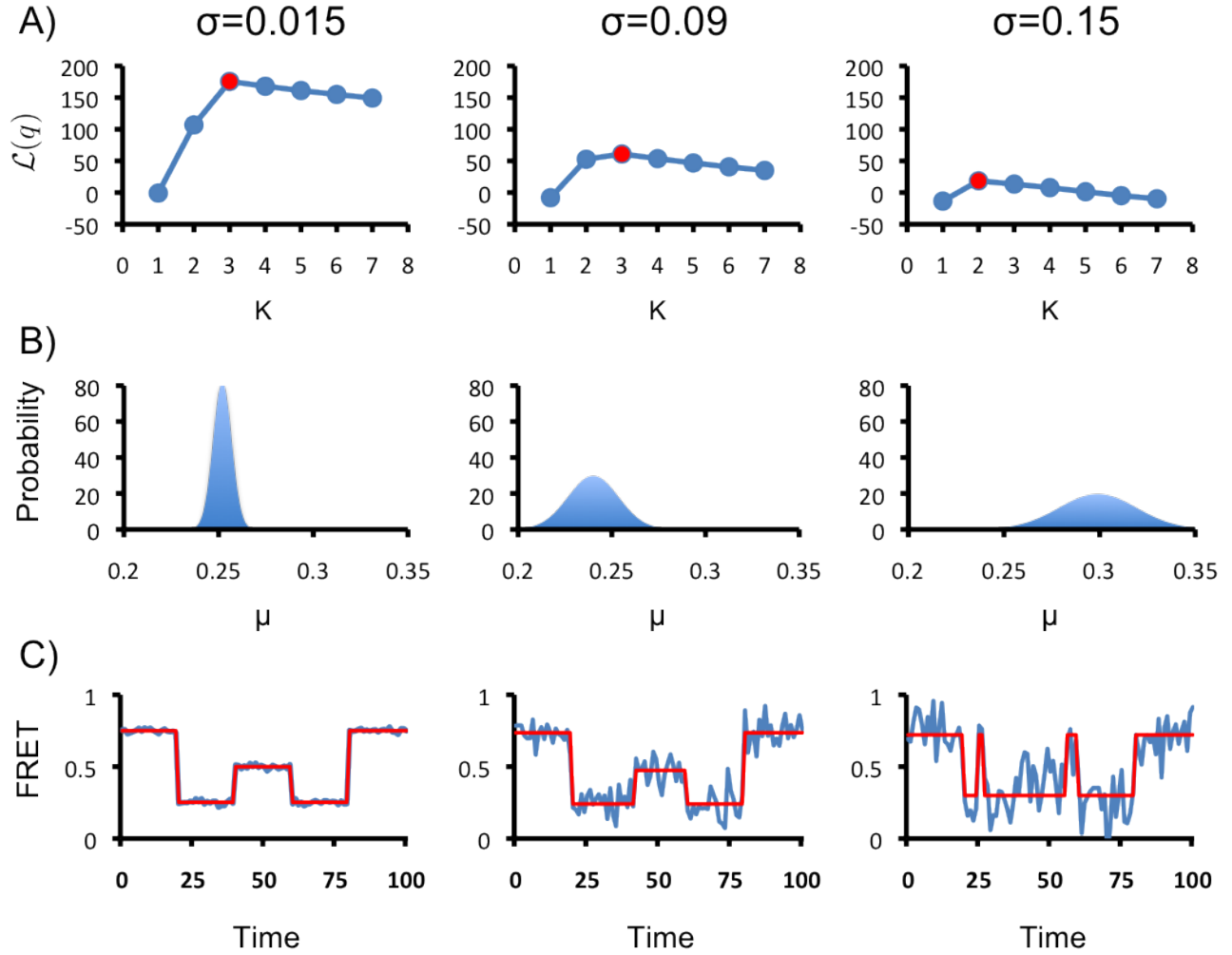


Figure 5: (A) Model selection using ME. Inference using $1 \leq K \leq 7$ hidden states was performed for each trace. The results with the highest $\mathcal{L}(q)$ are shown in red. (B) The posterior parameter distribution for the lowest-valued smFRET state inferred in each time series. The width of the posterior increases with the noise of the smFRET states, indicating lower confidence in the parameters learned from inference on noisier time series. (C) The idealized trajectories (red) inferred for each time series (blue) using the most probable parameters of the inference with the highest $\mathcal{L}(q)$.